学校编码：10384　　分类号_____密级_____

学号：23020101153040　　UDC_____

厦门大学

硕 士 学 位 论 文

选择性集成学习研究与应用

The Research and Application of Selective Ensemble Learning

邱诚

指导教师姓名：倪子伟　副教授

专 业 名 称：计算机软件与理论

论文提交日期：2013 年　6　月

论文答辩时间：2013 年　　月

学位授予日期：2013 年　　月

答辩委员会主席：

评　阅　人：_____

2013年　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：邱诚

2013 年 5 月 29 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（√）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：邱诚

2013 年 5 月 29 日

摘　要

集成学习利用现有简单学习算法共同解决一个问题，能够显著提高一个学习系统的泛化能力，对于未知的对象给出尽可能精确的估计。由于集成学习所具备的优势，近年来对其理论和算法的研究成为机器学习领域的热点之一。然而，为了获得满意的精度，集成大量的基分类器需要大量的存储空间并且使得预测速度明显下降，同时由于集成了大量冗余基分类器，影响了学习系统整体的泛化性。2002 年，周志华等人研究首先证实，并非所有基分类器参与集成的效果是最好，并且提出了“选择性集成”的概念。选择性集成学习是在已生成的多个基分类器的基础上，基于某种选择策略只从其中选择一部分用于构建集成分类器。本文在深入学习选择性集成研究以及与其相关理论的基础上，从混合模型方面进行了相应研究，主要完成以下工作：

（1）本文提出了将基于聚类的集成修剪和动态选择与循环集成框架相互结合的混合模型。首先基于 *K*-均值聚类算法的集成修剪算法剔除冗余的基分类器。然后，为了避免使用枚举法并且能够获得更好的集成性能，通过改进动态选择与循环集成框架，利用顺序选择策略对处理过的候选分类器集合进行集成学习。在多组实际数据集上进行对比实验，验证该模型处理二分类问题的有效性。

（2）本文将混合模型应用于音乐流派分类，考虑到音乐分类属于多分类问题，为了提高乐曲的识别精度，利用交叉验证对基分类器的参数进行初始化。实验表明混合模型适合处理音乐流派分类问题，并且通过交叉验证优化参数进一步提高性能。

（3） 通过问题转化方法衍生出应用于多标记分类的混合模型。在多标记数据集上进行对比实验，实验结果验证了混合模型在多标记问题上的可行性。

关键词： 选择性集成；聚类；动态选择；循环集成; 多标记学习

Abstract

Ensemble learning utilizes base learning algorithms to solve the problem, improves the generalization of learning system significantly and predicts unknown instance accurately. Because of the advantages of ensemble learning, the research on its theory and algorithm become one of hot spots in machine learning. However, integrating a number of base classifiers required to achieve a reasonable accuracy is enormously large and hence very space consuming. And the redundant classifiers have bad influence on the generalization of the whole system. In 2002, Zhou et al revealed that it might be better to ensemble many instead of all of base classifiers at hand and proposed the concept of selective ensemble. Selective ensemble is also a learning paradigm, trains a number of base classifiers and chooses some of them to ensemble by selective strategy. Based on the in-depth study of the research and theory on selective ensemble, this paper mainly focuses on hybrid model. The main contributions can be summarized as follows:

(1) The hybrid model combining of ensemble pruning based clustering and dynamic selection and circulating combination is presented. The first phase is ensemble pruning based on *K*-means clustering to eliminate redundant classifiers. The subsequent phase is the framework of dynamic selection and circulating combination with sequential search, which is aimed at improving the ensemble performance without the exhaustive enumeration. The comparative experiments on real datasets verify the validity of two-category classification of the hybrid model.

(2) The hybrid model is used to audio classification. Considering that audio classification is multi-class problem, cross validation is employed to initialize the parameters so as to improve the accuracy. The experiments illustrate that the hybrid model is suitable to musical genre classification and cross validation can improve the performance.

(3) The multi-label oriented model is derived from hybrid model by problem transformation. The comparative experiments on multi-label datasets verify its feasibility of multi-label classification.

Keywords: selective ensemble; clustering; dynamic selection; circulating combination; multi-label learning

目　录

Contents

第一章 绪论

“选择性集成”的概念是周志华等人在 2002 年首次提出的，所谓选择性集成学习，就是首先独立训练多个基分类器，然后通过一定的选择策略，从全部的候选分类器中选取部分分类器构成子集，这个子集中的分类器对于特定的数据具有较好的预测效果并且各个分类器彼此之间存在较大的差异性，通过集成子集中的分类器构建更好的预测系统。这个理论揭示了对于模式回归和分类问题，所有分类器都参与集成所得到的效果并非最好。该理论一经提出就在国内外的集成学习界引起了巨大轰动，并且在多个领域受到高度关注和应用。本章首先简述选择性集成学习的理论依据，然后论述了选择性集成学习的研究背景和意义，简要介绍了选择性集成学习在国内外的研究现状，最后介绍全文的内容安排。

1.1 选择性集成学习的理论依据

分类是集成学习的主要任务之一，它主要对未知样本的离散型类标记进行预测。本小节主要的任务是对选择性集成学习处理分类问题的有效性提供理论依据。为了更好地在下面的分析过程中进行讨论，假设面对二分类问题，未知样本的类标记由所有基分类器进行多数投票决定。需要说明一点，以下讨论内容可以推广到不同的集成方式以及处理多分类问题[1]。

假设 N 个基分类器函数 $f_1,...,f_N$，这些集成模拟的目标函数 $f:\Re^m \to \Omega$，其中 Ω 表示二分类的类标记集合，它只含有两种不同的类标记，逼近函数可以改写成 $f:\Re^m \to \{-1,+1\}$。假设训练集 D 含有 m 个样本，它们的期望类标记向量为 $[\omega_1,...,\omega_m]^{\mathrm{T}}$，其中 ω_j 表示第 j 个样本的期望标记。同时，第 i 个基分类器 f_i 输出 m 个样本的类标记向量为 $[f_{i1},...,f_{im}]^{\mathrm{T}}$，其中 f_{ij} 表示第 i 个基分类器对于第 j 个样本的输出类标记。ω_j 和 f_{ij} 分别满足 $\omega_j \in \{-1,+1\}(j=1,...,m)$，$f_{ij} \in \{-1,+1\}(i=1,...,N, j=1,...,m)$。显然，当第 i 个基分类器输出第 j 个样本

的期望类标记时，那么 $f_{ij}\omega_j = +1$；否则，$f_{ij}\omega_j = -1$。因此，第 i 个基分类器对于训练集 D 的泛化误差可以表示为：

$$E_i = \frac{1}{m}\sum_{j=1}^{m} Error(f_{ij}\omega_j) \qquad (1.1)$$

其中，函数 $Error$（x）为：

$$Error(x) = \begin{cases} 1 & x = -1 \\ 0.5 & x = 0 \\ 0 & x = +1 \end{cases} \qquad (1.2)$$

向量 $\boldsymbol{Sum}$ 为$[Sum_1,..., Sum_m]^{\mathrm{T}}$，其中 Sum_j 表示全部的基分类器对第 j 个样本输出情况，它可以定义为：

$$Sum_j = \sum_{i=1}^{N} f_{ij} \qquad (1.3)$$

因此，全部基分类器集成输出第 j 个样本类标记为：

$$\hat{f}_j = Sgn(Sum_j) \qquad (1.4)$$

其中函数 Sgn（x）为：

$$Sgn(x) = \begin{cases} +1 & if \quad x > 0 \\ 0 & if \quad x = 0 \\ -1 & if \quad x < 0 \end{cases} \qquad (1.5)$$

显然，$\hat{f}_j \in \{-1,0,+1\}(j = 1,..., m)$。如果集成分类器输出结果为第 j 个样本的期望标记，那么 $\hat{f}_j\omega_j = +1$；否则，$\hat{f}_j\omega_j = -1$；如果正反例投票个数相同，那么 $\hat{f}_j\omega_j = 0$。因此，集成分类器对于训练集 D 的泛化误差可以定义为：

$$\hat{E} = \frac{1}{m}\sum_{j=1}^{m} Error(\hat{f}_j\omega_j) \qquad (1.6)$$

下面设想从全部基分类器中剔除第 k 个基分类器，那么新的集成分类器对于第 j 个样本的输出变为：

$$\hat{f}_j' = Sgn(Sum_j - f_{kj}) \qquad (1.7)$$

而且集成分类器对于训练集 D 的泛化误差变为：